U.S. Patent Application based on PCT/EP98/05793

Summary of DE 195 30 646 C1

DE 195 30 646 C1 discloses a learning method for a recurrent neural network. By the application of this method, the returns of recurrent neural networks are splitted. This provides additional inputs receiving training data. The training data are obtained on the basis of error distributions and the application of a Monte Carlo method.

DE 195 30 646 C1 represents technological background with regard to the use of neural networks. It does not disclose a method for detecting the modes of a dynamic system with a drift segmentation model as claimed in the above U.S. patent application.

(19) BUNDESREPUBLIK DEUTSCHLAND

DEUTSCHES PATENTAMT

(12) **Patentschrift**

(10) **DE 195 30 646 C 1**

(51) Int. Cl.⁶:
G 06 F 15/18
G 05 B 13/02

(21) Aktenzeichen: 195 30 646.5-53
(22) Anmeldetag: 21. 8. 95
(43) Offenlegungstag: —
(45) Veröffentlichungstag der Patenterteilung: 17. 10. 96

Innerhalb von 3 Monaten nach Veröffentlichung der Erteilung kann Einspruch erhoben werden

(73) Patentinhaber:
Siemens AG, 80333 München, DE

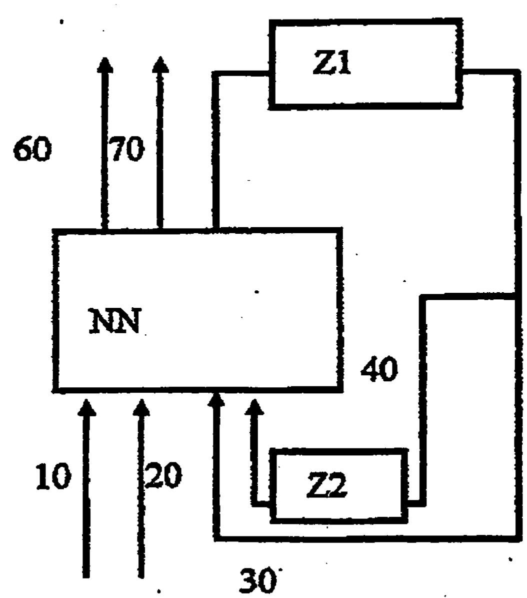
(72) Erfinder:
Tresp, Volker, Dr., 80539 München, DE

(56) Für die Beurteilung der Patentfähigkeit in Betracht gezogene Druckschriften:

US 51 82 794
TRESP, Volker, et.al.: Training Neural Networks with Deficient Data. In: Cowan, J.D.et.al.(Hrsg.):Advances in Neural Information Processing Systems 6, Morgan Kaufman, 1994, S. 128-135;
JP 06-301 663 A,. In: Patent Abstracts of Japan (CD-ROM);

(54) Lernverfahren für ein rekurrentes neuronales Netz

(57) Bei Anwendung des erfindungsgemäßen Verfahrens werden die Rückführungen rekurrenter neuronaler Netze aufgetrennt. Für die hierdurch gewonnenen zusätzlichen Eingänge am Netz werden Trainingsdaten künstlich bereitgestellt, indem anhand der bekannten Meßwerte aus der zu Trainingszwecken zugeführten Zeitreihe und deren bekannter oder vorgegebener Fehlerverteilungsdichte für die zusätzlichen Werte Fehlerverteilungsdichten berechnet werden. Hierzu sind gemäß der Monte-Carlo-Methode aus dieser Fehlerverteilungsdichte Proben zu ziehen. Mit denen wird je ein vorherzusagender Wert der Zeitreihe berechnet und deren Mittelwert wird für einen vorherzusagenden Wert eingesetzt. Einsatzgebiete der Erfindung liegen auf allen bekannten Einsatzgebieten neuronaler Netze.



DE 195 30 646 C1

Beschreibung

Die Erfindung bezieht sich auf ein Lernverfahren zur neuronalen Modellierung von dynamischen Prozessen.

Neuronale Netze finden in die vielfältigsten technischen Gebiete Eingang. Überall dort, wo es gilt, aus komplexen technischen Zusammenhängen und aus unzureichenden Informationen Entscheidungen abzuleiten, erweisen sich neuronale Netze als besonders geeignet. Zur Bildung einer oder mehrerer Ausgangsgrößen werden dem neuronalen Netz beispielsweise eine oder mehrere Eingangsgrößen zugeführt. Hierzu wird ein solches Netz zunächst für den speziellen Einsatzfall trainiert, anschließend generalisiert und danach wird es mit einem anderen Datensatz als den Trainingsdaten validiert. Neuronale Netze erweisen sich für viele Einsatzfälle als besonders geeignet, da sie universell trainierbar sind.

Ein häufig auftretendes Problem im Zusammenhang mit dem Einsatz von neuronalen Netzen besteht allerdings darin, daß häufig die Eingangsdaten zum Training, oder beim Betrieb des Netzes nicht vollständig sind. Dieser Sachverhalt und auch die Tatsache, daß die Meßwerte für den Aufbau einer Zeitreihe, welche dem neuronalen Netz zugeführt wird, häufig ungenau oder verrauscht sind, bewirken, daß teilweise schlechte Lernergebnisse der Netze erzielt werden. Bei rekurrenten neuronalen Netzen tritt insbesondere das Problem auf, daß wegen der auf Eingänge rückgekoppelten Ausgänge beim Training nicht alle Eingänge mit Trainingsdaten beschickt werden können. Bislang gibt es keine Methoden, welche dieser besonderen Problematik Rechnung tragen.

Aus der Veröffentlichung von Tresp, Volker et al.: "Training Neural Networks with Deficient Data". In: Cowan, J.D. et al. (Hrsg.): Advances in Neural Information Processing Systems 6, Morgan Kaufman, 1994, S. 128—135, sind die Grundlagen zum Training neuronaler Netze mit Hilfe unvollständiger Daten bekannt. Vor allen Dingen werden dabei Lösungen bevorzugt, die eine gewichtete Integration über die unbekannten oder unbestimmten Eingänge erfordern.

Aus der US-Patentschrift US 5 182 794 ist ein Lernsystem für rekurrente neuronale Netze bekannt. Dieses Lernverfahren betrifft neuronale Netze, welche versteckte Eingangs- und Ausgangsneuronen haben und kalkuliert Wichtungsfehler über eine beschränkte Anzahl von Fortpflanzungen über das Netzwerk. Mit diesem Verfahren können gängige Lernmengen, welche für vorwärtspropagierende Netzwerke verwendet werden, auch für rekurrente Netzwerke Verwendung finden.

Aus der japanischen Veröffentlichung JP 06-301 663 A. In: Patent Abstracts of Japan ist ein Lernsystem für seriell verbundene rekurrente neuronale Netze bekannt. Dabei wird der aktive Ausgangswert eines Ausgangsneurons 4 als Lernsignal für einen Ausgangszustand verwendet, und in einem fortgeschrittenen Stadium wird beim Updaten der Verbindung der Ausgangswert des Neurons 4 und der Eingangswert des Lernsignals mit einer vorbeschriebenen Gewichtungsfunktion gewichtet.

Die der Erfindung zugrundeliegende Aufgabe besteht deshalb darin, ein Lernverfahren anzugeben, mit dem der Lernvorgang beim Training eines rekurrenten neuronalen Netzes verbessert werden kann.

Diese Aufgabe wird gemäß den Merkmalen des Patentanspruchs 1 gelöst.

Weiterbildungen der Erfindung ergeben sich aus den abhängigen Ansprüchen.

Ein besonderer Vorteil des erfindungsgemäßen Verfahrens besteht darin, daß hier erstmals ein Verfahren vorgestellt wird, mit welchem mehr Eingänge für das Training rekurrenter neuronaler Netzwerke zur Verfügung stehen. Durch das Verfahren werden nämlich die Rückführungen der rekurrenten Netze aufgetrennt und diese aufgetrennten Verbindungen werden als Eingänge für künstlich erzeugte Werte benutzt. Vorteilhafterweise stellt das erfindungsgemäße Verfahren hochwertige künstlich erzeugte Werte für die Zeitreihe zur Verfügung, welche mit Hilfe der Fehlerverteilungsdichte und der restlichen bekannten Werte der Zeitreihe, mit welchem das Netz trainiert werden soll, berechnet werden.

Besonders vorteilhaft kann das erfindungsgemäße Verfahren auch angewendet werden, wenn mehrere nebeneinander liegende Werte der Zeitreihe unbekannt sind. Dies ist der Fall, falls im rekurrenten Netzwerk mehrere Verbindungen aufgetrennt werden. Zur Berechnung dieser Werte wird vorzugsweise ein Iterationsverfahren angegeben, bei welchem ähnlich vorgegangen wird, wie bei der Berechnung eines Einzelwertes, jedoch wird dann dieser Wert als bekannt vorausgesetzt und der weitere unbekannte Wert mit diesem berechnet. Die Iteration kann solange durchgeführt werden, bis eine hinreichende Genauigkeit der zu ermittelnden Werte gefunden wurde, oder bis das Netz mit dem gewünschten Lernerfolg trainiert wurde.

Ein besonderer Vorteil des erfindungsgemäßen Verfahrens besteht darin, daß ausgenutzt wird, daß die fehlenden Werte oder die verrauschten Werte, welche den neuronalen Netz zugeführt werden sollen, Bestandteil einer Abfolge von Werten in der Zeitreihe sind. Vorteilhaft kann die bekannte Fehlerverteilungswahrscheinlichkeit der restlichen Werte dazu benutzt werden, um für den fehlenden Wert nach dem erfindungsgemäßen Verfahren eine erwartete Fehlerverteilung und damit den erwarteten Wert berechnen zu können.

Günstigerweise können nach dem erfindungsgemäßen Verfahren auch fehlende Werte, die in der Zeitreihe benachbart sind, bestimmt werden. Dafür ist ein iterativer Vorgang vorgesehen, der einmal den einen Wert berechnet und danach den anderen mit den aus dem einen Wert gewonnenen Daten ermittelt. Günstigerweise kann dieser Iterationsvorgang auch mehrfach durchgeführt werden, damit eine hinreichende Genauigkeit der zu bestimmenden Werte gewährleistet ist.

Besonders kann nach dem erfindungsgemäßen Verfahren auch ein neuronales Netz trainiert werden, welches die Zeitreihe nachbilden soll, da dabei vorteilhaft die Lernschrittweite auf die Zahl der gezogenen Monte Carlo Proben bezogen wird.

Günstigerweise wird für das Fehlerverhalten der Eingangsgrößen der Zeitreihe eine Gaußverteilung angenommen oder festgelegt, da dies eine Verteilung ist, welche praxisnahen Werten weitestgehend entspricht.

Vorteilhaft werden einfache numerische Verfahren beim erfindungsgemäßen Verfahren eingesetzt, um fehlende Werte bestimmen zu können bzw. um einen künftigen Wert in der Zeitreihe mit fehlenden Werten

vorhersagen zu können.

Vorteilhaft werden nach dem erfindungsgemäßen Verfahren einfache mathematische Methoden angegeben, um Meßwerte entsprechend aufbereiten zu können und damit das neuronale Netz trainieren zu können.

Besonders vorteilhaft werden Methoden angegeben, um verrauschte Meßwerte aufzubereiten, bzw. um Meßwerte aufzubereiten, welche eine bekannte und eine unbekannte Rauschkomponente enthalten, um damit das neuronale Netz auf einfache Weise und effizient trainieren zu können.

Im folgenden wird die Erfindung anhand von Figuren weiter erläutert.

Fig. 1 zeigt ein rekurrentes neuronales Netz

Fig. 2 gibt ein Beispiel des erfindungsgemäßen Verfahrens an

Fig. 3 zeigt eine Zeitreihe

Fig. 4 zeigt eine Zeitreihe und ein Systemverhalten

Fig. 1 gibt ein Beispiel eines rekurrenten neuronalen Netzwerkes NN an. Beispielsweise wird dieses Netz NN mit einem Systemverhalten SY trainiert, welches in Fig. 4 näher dargestellt ist. Am Netz sind freie Eingänge 10, 20 und freie Ausgänge 60, 70 vorgesehen. Bei bisher angewandten Trainingsverfahren für solche rekurrenten Netze stehen lediglich die Eingänge 10 und 20 zur Zufuhr von Trainingsdaten bereit.

Weiter sind am Netz Rückführungen 30 und 40 vorgesehen, welche beim rekurrenten Netz ein quasi Gedächtnis realisieren. In Fig. 1 ist weiterhin eine Schnittlinie 50 angedeutet, entlang der nach dem erfindungsgemäßen Verfahren die Rückführungen bzw. die Leitungen zwischen den Ausgängen und Eingängen 30 und 40 aufgetrennt werden können. Vorteilhaft wird so durch das erfindungsgemäße Verfahren der Effekt erzielt, daß beim Training des rekurrenten Netzes auch die Eingänge 30 und 40, welche sonst geschlossen wären, für den Trainingsvorgang zur Verfügung stehen. Dadurch, daß beim Trainieren des Netzes mehrere Eingänge verwendet werden, wird ein schnellerer Trainingserfolg erreicht. Bei der Bereitstellung der Trainingsdaten für diese zusätzlichen Eingänge wird erfindungsgemäß davon ausgegangen, daß das Netz mit einer Zeitreihe von zeitlich nacheinander anfallenden Meßwerten eines technischen Systems beschickt werden soll und dieses nach dem Training nachbilden soll. Da, so die erfinderische Idee, die beiden zusätzlichen Eingänge 30 und 40 ebenfalls Bestandteil der Zeitreihe von Meßwerten sind, kann über die restlichen Meßwerte und deren Fehlerverteilungsdichte der Wert errechnet werden. Dies geschieht über das erfindungsgemäße Verfahren mittels der angenommenen Fehlerverteilungsdichte für die Eingänge 30 und 40 und der Ziehung von Monte-Carlo-Proben gemäß dieser Fehlerverteilungsdichte und Mittelwertbildung der gezogenen Proben. So ergeben sich zusätzliche Trainingseingangswerte gemäß dem erfindungsgemäßen Verfahren. Die genaueren Details dieser Berechnungsmethode mit dem erfindungsgemäßen Verfahren sind bei den Erläuterungen der Fig. 3 und 4 dargestellt.

In Fig. 2 wird der erfindungsgemäße Ansatz nochmals anhand eines Beispieles verdeutlicht. In Fig. 2 ist dasselbe rekurrente neuronale Netz wie in Fig. 1 dargestellt. Bei der erfindungsgemäßen Lösungsvariante wird davon ausgegangen, obwohl das in Realität nicht der Fall ist, daß die beiden durch Auftrennung der Leitung 50 entstehenden Eingänge 30 und 40 mit Werten einer Zeitreihe beschickt werden, welche sich aus den Eingangswerten der Eingänge 10 und 20 errechnen lassen. Der Eingang 30 wird also beispielsweise mit dem Wert M und der Fehlerverteilung ε1 aus der Fig. 4 beschickt, welcher in der Zeitreihe die Position y_{t-2} innehat. Der Eingang 40 wird dann beispielsweise mit dem Wert L aus Fig. 4 und dessen Fehlerverteilungsdichte ε2 beschickt. Dieser nimmt in der Zeitreihe die Position y_{t-3} ein. Mit Z1 und Z2 sind Zeitglieder bezeichnet, welche der unterschiedlichen Verzögerung der einzelnen Werte innerhalb der Zeitreihe entsprechen. Z1 entspräche in diesem Fall einer Zeitverzögerung von y_{t-1} bis y_{t-2} und Z2 entspräche in diesem Fall an der Leitung 40 einer Zeitverzögerung zwischen y_{t-2} und y_{t-3}. Mit der erfindungsgemäßen Methode wird ein Verfahren angegeben, mit welchem fehlende Werte einer Zeitreihe mit hoher Genauigkeit über deren Fehlerverteilungsdichte und Ziehen von Monte-Carlo-Proben mit anschließender Mittelwertbildung nachgebildet werden können. Der Grundansatz der Erfindung macht sich diese Idee zunutze, indem die künstlich erzeugten Eingänge von rekurrenten neuronalen Netzen mit künstlich erzeugten, durch die erfindungsgemäße Methode berechneten Werten, welche aber sehr genau sind, beschickt werden. Hierdurch wird vorteilhaft eine bessere Trainingsleistung, welche sich in genaueren Gewichten an den einzelnen Neuronen widerspiegelt, erzielt. Es ist nicht erforderlich, zur Durchführung des erfindungsgemäßen Verfahrens alle Leitungen, d. h. Rückführungen des neuronalen Netzes, aufzutrennen, vielmehr können auch fallweise lediglich einzelne Leitungen aufgetrennt werden, da so erheblich weniger Rechenaufwand bei der Ziehung der Monte-Carlo-Proben und der Berechnung des Wertes, über die Zugrundelegung der Fehlerverteilungsdichte der restlichen Zeitreihe, getrieben werden muß. Bei komplexeren Vorgängen kann es auch sinnvoll sein, mehrere Rückführungsleitungen des neuronalen Netzes, als hier in Fig. 1 und 2 dargestellt sind, aufzutrennen und auf komplizierte Weise die hierfür erforderlichen Eingangswerte zu bestimmen. Vorteilhaft wird mit dem erfindungsgemäßen Verfahren ein Iterationsverfahren angegeben, welches auf Basis der erfindungsgemäßen Methode eine Berechnung auch mehrerer Werte ermöglicht. Fallweise wird der Fachmann, welcher die Erfindung ausführt, entscheiden können, welche Methode zur Durchführung der Erfindung für ihn die günstigste ist. Er wird dabei abwägen, ob er hinreichend Rechenkapazität zur Verfügung hat, um das Netz genau modellieren zu können, oder ob es ihm wichtiger ist, einen schnellen Erfolg beim Trainieren des neuronalen Netzes zu erreichen.

Fig. 3 zeigt eine Zeitreihe von Meßwerten, welche beispielsweise einem neuronalen Netz zugeführt werden können. Gemäß ihrer zeitlichen Abfolge werden diese Meßwerte beispielsweise von einem technischen System erfaßt und gemäß ihrer zeitlichen Abfolge mit y_t bis y_{t-6} bezeichnet. Die dargestellten Pfeile zwischen den einzelnen Kästchen symbolisieren die Abhängigkeiten der verschiedenen Werte untereinander. Beispielsweise wird in Fig. 3 davon ausgegangen, daß der Wert y_{t-2} fehlt. Die im Markov blanket relevanten Werte, als benachbarte Werte dieses fehlenden Meßwertes, sind y_{t-4}, y_{t-3}, y_{t-1} und y_t. Ein solch fehlender Meßwert in einer Zeitreihe kann beispielsweise dadurch entstehen, daß zum fraglichen Zeitpunkt das Meßgerät, zur Werteaufnahme nicht funktionierte, oder daß es zwischen einzelnen gemessenen Werten günstig erscheint, um das

neuronale Netz besser zu trainieren, diesem einen weiteren Wert zuzuführen, der folglich noch zu bestimmen ist, also noch dem erfindungsgemäßen Verfahren erzeugt werden soll. Beispielsweise wird in Fig. 3 weiter davon ausgegangen, daß der Wert y_{t-3} fehlt. Die im Markov blanket relevanten Werte, als benachbarte Werte dieses fehlenden Meßwertes, sind y_{t-5}, y_{t-4}, y_{t-2} und y_{1-4}. Beim vorgestellten erfindungsgemäßen Verfahren entstehen diese fehlenden Meßwerte jedoch durch Anwendung des erfindungsgemäßen Verfahrens auf rekurrente neuronale Netze. Indem die Rückkopplungsleitungen 30 und 40 aufgetrennt werden, ergeben sich hier beispielsweise zwei zusätzliche Eingänge, welche bei Anwendung des erfindungsgemäßen Verfahrens zum Training zusätzlich mit Lerndaten beschickt werden können. Dies hat nach der erfinderischen Idee zur Folge, daß das Netz mit besser zutreffenden Gewichtungsfaktoren ausgestattet wird. Dies ist der Fall, obwohl die Werte, welche über diese Eingänge zugeführt werden auf die erfindungsgemäße Art quasi künstlich erzeugt werden. Dabei ist allgemein zu bemerken, daß die Anzahl der aufzutrennenden Rückführleitungen lediglich von der Ausprägung des rekurrenten neuronalen Netzes abhängt und nicht durch das erfindungsgemäße Verfahren beschränkt wird. Es kann fallweise auch günstig sein, nicht alle dieser Eingangs-Ausgangs-Verbindungen aufzutrennen, weil trotz des dabei zu betreibenden geringeren Rechenaufwandes eine hinreichende Lernleistung des Netzes erzielbar ist.

Fig. 4 zeigt die Zeitreihe aus Fig. 3 in Verbindung mit einem rekurrenten neuronalen Netz NN. Es ist zu erkennen, daß y eine zeitabhängige Variable darstellt, welche das Systemverhalten SY eines technischen Systems repräsentiert. Wie erkannt werden kann, entsprechen die Werte y_t bis y_{t-6} Meßwerten, welche dem Systemverlauf SY entnommen werden. Durch die gestrichelten Pfeile zu den jeweiligen Zeitpunkten ist symbolisiert, daß diese Meßwerte dem neuronalen Netz NN beim Training zugeführt werden sollen.

Wie auch in Fig. 3 ist der fragliche Meßwert M für den Zeitpunkt y_{t-2} nicht vorhanden. Für diesen Meßwert M ist seine Wahrscheinlichkeitsdichte ε1 angegeben. Diese Wahrscheinlichkeitsdichte ε kann beispielsweise nach dem erfindungsgemäßen Verfahren aus einer vorgegebenen bekannten Fehlerverteilungsdichte der übrigen bekannten Meßwerte rückgerechnet werden. Insbesondere wird dabei ausgenutzt, daß sich der fehlende Meßwert zwischen zwei bekannten Meßwerten befinden muß und damit auch dessen Fehler durch die Fehler der benachbarten und der restlichen Meßwerte der Zeitreihe begrenzt wird. Die zugrundeliegende Zeitreihe läßt sich wie folgt beschreiben:

$$y_t = f(y_{t-1}, y_{t-2}, \ldots, y_{t-N}) + \varepsilon_t \quad (1)$$

Dabei ist f entweder bekannt oder wird hinreichend durch ein neuronales Netz modelliert. ε_t bedeutet dabei einen additiven unkorrelierten Fehler mit zeitlichem Mittelwert 0. Dieser Fehler weist dabei — und das ist für das erfindungsgemäße Verfahren essentiell — eine bekannte oder vorgegebene Wahrscheinlichkeitsdichte $P_\varepsilon(\varepsilon)$ auf und versinnbildlicht typischerweise die unmodellierte Dynamik der Zeitreihe. Beispielsweise soll für eine solche Zeitreihe, die nach dem erfindungsgemäßen Verfahren komplettiert werden soll, ein zukünftiger Wert vorhergesagt werden. Dabei ist zu beachten, daß zukünftige Werte relativ zu der momentanen gewählten Zeitposition zu verstehen sind. Das heißt für einen Zeitpunkt y_{t-5} ist der Zeitpunkt y_{t-4} ein zukünftiger Wert. Unter diesen Voraussetzungen läßt sich die bedingte Wahrscheinlichkeitsdichte für einen vorherzusagenden Wert der Zeitreihe wie folgt beschreiben.

$$P(y_t|y_{t-1}, y_{t-2}, \ldots, y_{t-N}) = P_\varepsilon(y - f(y_{t-1}, y_{t-2}, \ldots, y_{t-N})) \quad (2)$$

Wie bereits erwähnt muß die Fehlerverteilungsdichte bekannt sein. Diese Verteilungsdichte kann entweder anhand des Systemverhaltens und bekannter anderer äußerer Größen ermittelt oder vorgegeben werden. Eine typische Fehlerverteilung, die in der Praxis auftritt ist die Gaußverteilung. Mit einer solchen angenommenen Gauß'schen Fehlerverteilung läßt sich die bedingte Wahrscheinlichkeitsdichte wie folgt beschreiben:

$$P(y_t|y_{t-1}, y_{t-2}, \ldots, y_{t-N}) = G(y_t; f(y_{t-1}, \ldots, y_{t-N}), \sigma^2) \quad (3)$$

Darin bedeutet $G(x; c, \sigma^2)$ die Notation für eine normale Dichte, die bei x bestimmt wird mit einem Zentrum C und einer Varianz σ^2. Geht man davon aus, daß das zu beschreibende System in Form einer Folge von Werten auf einer Zeitachse dargestellt wird, so kann man die einzelnen Werte von y_t auch als Zufallsvariable in einem probabilistischen Netzwerk auffassen. Dabei liegt der Erfindung das Problem zugrunde, einen Wert der Zeitreihe vorherzusagen, indem die vorhandene Information aus den restlichen Werten möglichst vollständig verwendet wird. Unter Voraussetzung der Annahmen, die zuvor gemacht wurden, läßt sich die gesamte Wahrscheinlichkeitsdichte der Zeitreihe wie folgt beschreiben:

$$P(y_1, y_2, \ldots, y_t) = P(y_1, \ldots, y_N) \prod_{l=N+1}^{t} P(y_l|y_{l-1}, \ldots, y_{l-N}) \quad (4)$$

Dabei wird davon ausgegangen, daß y_{t-k} mit $k \leq N$ der fehlende Wert ist. Vorausgesetzt, daß $y^u = \{y_{t-k}\}$ und $y^m = \{y_{t-1}, \ldots, y_{t-k-N}\}/\{y_{t-k}\}$ gilt, kann der erwartete Wert der in der Zeitreihe vorherzusagen ist wie folgt beschrieben werden:

$$E(y_t|M_{t-1}) = \int f(y_{t-1}, \ldots, y_{t-k}, \ldots, y_{t-N}) P(y^u|y^m) dy^u \quad (5)$$

Dabei gelten folgende Voraussetzungen:

M_{t-1} steht für alle Messungen bis zum Zeitpunkt $t-1$. Die voranstehende Gleichung ist die grundlegende Gleichung für die Vorhersage mit fehlenden Daten. Dabei ist besonders zu beachten, daß die Unbekannte y_{t-k} nicht nur von den Werten der Zeitreihe vor dem Zeitpunkt $t-k$ abhängt, sondern auch von den Messungen nach $t-k$. Der Grund besteht darin, daß die Variablen in $y^m \cup y_t$ ein minimales Markov blanket von y_{t-k} formen. Dieses minimale Markov blanket besteht aus den direkten Vorfahren und den direkten Nachfahren einer Variable und allen direkten Vorfahren von Variablen des direkten Nachfolgers. Im betrachteten Beispiel in Fig. 4 sind die direkten Nachfahren $y_t \ldots y_{t-k+1}$. Die direkten Vorfahren sind:

$y_{t-k-1} \ldots y_{t-k-N}$

und die direkten Eltern des Nachfolgers der Variablen sind:

$y_{t-1} \ldots y_{t-k-N+1}$.

Aus den theoretischen Grundlagen ist bekannt, daß eine Variable unabhängig von einer anderen Variablen dieses Netzwerkes ist, wenn die Variablen innerhalb des Markov blankets bekannt sind. Deshalb wird die benötigte bedingte Dichte aus Gleichung (5) wie folgt bestimmt:

$$P(y^u|y^m) \propto P(y_{t-1}|y_{t-2}, \ldots, y_{t-k}, \ldots, y_{t-1-N}) \times P(y_{t-2}|y_{t-3}, \ldots, y_{t-k}, \ldots, y_{t-2-N}) \ldots P(y_{t-k}|y_{t-k-1}, \ldots, y_{t-k-N}) \quad (5b)$$

Der hier beschriebene Fall eines fehlenden Meßwertes kann auch ohne Beschränkung der Erfindung auf mehrere nebeneinander liegende fehlende Meßwerte ausgedehnt werden. Falls dies der Fall ist, muß beispielsweise nach dem erfindungsgemäßen Verfahren zunächst der eine Wert anhand seiner Nachbarn und Eltern und Vorfahren bestimmt werden und mit diesem bestimmten Wert dann der weitere Wert ermittelt werden. Dies kann solange hin- und hergehen bis eine hinreichende Genauigkeit erreicht werden. Für diesen Fall gilt:

$$y^u \subseteq \{y_{t-1}, y_{t-2}, \ldots, y_{t-n}\} \quad (5c)$$

Für alle fehlenden Werte der Zeitreihe zwischen dem Zeitpunkt $t-1$ und $t-N$, wobei weiterhin gilt:

$$y^m \subseteq \{y_{t-1}, y_{t-2}, \ldots, y_t\} \quad (5d),$$

welche die Zahl aller Meßwerte bis zum Zeitpunkt $t-1$ repräsentiert. Auch gilt

$$P(y^u|y^m) \sim P(y_{t-1}, \ldots, y_2, y_1) \quad (5e),$$

wobei die rechte Seite in (5e) aus Gleichung (4) erhalten wird. Im allgemeinen sind diese Integrale in den voranstehenden Gleichungen für die Funktion f(), falls dies eine nichtlineare Funktion ist, nicht analytisch lösbar. Details für die numerische Lösung mit Hilfe der Ziehung von Monte-Carlo-Proben werden im Zusammenhang mit Fig. 3 angegeben. Für den Fall, daß ein weiterer Meßwert der Zeitreihe nachgebildet werden soll, wie dies bei zwei aufgetrennten Rückkopplungsverbindungen der Fall ist, sieht das erfindungsgemäße Verfahren eine iterative Approximation der fehlenden Werte vor. Beispielsweise ist für das Training des Netzes zusätzlich der Wert L für den Zeitpunkt y_{t-3} nachzubilden. Für diesen Meßwert M ist seine Wahrscheinlichkeitsdichte ε2 angegeben. Diese Wahrscheinlichkeitsdichte ε2 kann beispielsweise nach dem erfindungsgemäßen Verfahren aus einer vorgegebenen bekannten Fehlerverteilungsdichte der übrigen bekannten Meßwerte rückgerechnet werden. Für die Berechnung von zwei solchen Werten L und M wird zunächst L beispielsweise als bekannt vorausgesetzt oder geschätzt. Daraus wird M berechnet und mit dem bekannten Wert M wird anschließend erneut L bestimmt. Dieser Iterationsvorgang läuft vorzugsweise so lange ab, bis eine hinreichende Genauigkeit der Werte gegeben ist, oder bis das Netz genau genug trainiert ist.

Für den Fall, daß $y_1, \ldots y_t$ mögliche Werte der Zeitreihe darstellen, sollen $y^m \subseteq \{y_1, \ldots, y_t\}$ alle Meßwerte bezeichnen und $y^u = \{y_1, \ldots, y_t\} \backslash y^m$ alle unbekannten Werte bezeichnen. Das neuronale Netz NN, welches die Funktion f modelliert, werde beispielsweise mit einem Satz von Gewichten w parametrisiert. Dann gilt:

$$f(y_{t-1}, \ldots, y_{t-N}) \approx NN_w(y_{t-1}, \ldots, y_{t-N})$$

Die logarithmische Wahrscheinlichkeitsfunktion lautet dann:

$$L = \log \int P^M(y_t, y_{t-1}, \ldots, y_2, y_1) dy^u$$

wobei dann die gemeinsame Wahrscheinlichkeitsdichte sich zu

$$P^M(y_t, y_{t-1}, \ldots, y_2, y_1) = P^M(y_N, \ldots, y_1) \prod_{l=N+1}^{t} P^M(y_l | y_{l-1}, \ldots, y_{l-N}) \quad (6)$$

approximiert und für das neuronale Netz folgender Zusammenhang für die Berechnung der Fehlerverteilungs-

dichte gilt:

$$P^M(y_t|y_{t-1}, y_{t-2}, \ldots, y_{t-N}) = P_\varepsilon(y_t - NN_w(y_{t-1}, y_{t-2}, \ldots, y_{t-N})) \quad (7)$$

Für das Lernen mit Hilfe von Backpropagation, oder anderer Gradienten-basierter Lernalgorithmen wird nun noch der Gradient der logarithmischen Wahrscheinlichkeitsfunktion benötigt, welcher sich zu:

$$\frac{\partial L}{\partial w} = \sum_{t=N+1}^{t} \int \frac{\partial \log P^M(y_t|y_{t-1}, \ldots, y_{t-N})}{\partial w} P^M(y^{u(t)} \mid y^m) dy^{u(t)} \qquad (8)$$

ergibt. Es ist anzumerken, daß hierbei von bekannten Ausgangsbedingungen für $y_1, \ldots, y_N$ ausgegangen wird. Für den Fall, daß eine Gaußverteilung für die Fehlerverteilung vorliegt, ergibt sich daraus:

$$\frac{\partial L}{\partial w} \propto \sum_{t=N+1}^{t} \int (y_t - NN_w(y_{t-1}, \ldots, y_{t-N})) \frac{\partial NN_w(y_{t-1}, \ldots, y_{t-N})}{\partial w} P^M(y^{u(t)} \mid y^m) dy^{u(t)} \quad (8a)$$

wobei $y^{u(t)} = y^u \cap \{y_t, \ldots y_{t-N}\}$ die fehlenden Werte für die Eingänge des Netzwerkes darstellen und (8a) zeigt, daß, falls alle $y_t \ldots y_{t-N}$ bekannt sind, das Integral verschwindet.

Falls die Meßwerte von einem zusätzlichen aber bekannten Rauschen überlagert werden ergeben sich die folgenden Zusammenhänge. Beispielsweise gilt wieder:

$$y_t = f(y_{t-1}, y_{t-2}, \ldots, y_{t-N}) + \varepsilon_t$$

In dieser Variante der Erfindung soll jedoch kein direkter Zugriff auf y_t bestehen. Anstattdessen wird die Zeitreihe

$$z_t = y_t + \delta_t$$

gemessen. Darin bedeutet δ_t ein unabhängiges Rauschen mit Mittelwert Null. Unter der Voraussetzung, daß $z = \{z_1 \ldots z_{t-1}\}$ und $y = \{y_1 \ldots y_t\}$ gelten, ergibt sich die Gesamtwahrscheinlichkeitsdichte zu:

$$P(y, z) = P(y_N, \ldots, y_1) \prod_{l=N+1}^{t} P(y_l|y_{l-1}, \ldots, y_{l-N}) \prod_{l=1}^{t} P(z_l|y_l) \qquad (8b)$$

Damit läßt sich die Rechenvorschrift für den erwarteten nächsten Wert der Zeitreihe angeben:

$$E(y_t|z = \int f(y_{t-1}, \ldots, y_{t-N}) P(y_{t-1}, \ldots, y_{t-N}|z) dy_{t-1} \ldots dy_{t-N} \quad (9)$$

Ebenso kann der Gradient der Wahrscheinlichkeitsfunktion für das Training berechnet werden. Für den Fall, daß eine Gaußverteilung des Rauschens mit

$$z = \{z_1 \ldots z_t\}$$

vorliegt, ergibt sich:

$$\frac{\partial L}{\partial w} \propto \sum_{l=N+1}^{t} \int (y_l - NN_w(y_{l-1}, \ldots, y_{l-N})) \frac{\partial NN_w(y_{l-1}, \ldots, y_{l-N})}{\partial w} \times P^M(y_1, \ldots, y_{l-N}|z) dy_1 \ldots dy_{l-N} \qquad (9a)$$

Dem neuronalen Netz werden bei einer Variante des erfindungsgemäßen Verfahrens beispielsweise Werte zugeführt, die verrauscht oder nicht genau bestimmbar sind. Durch die Approximation der Gewichte im neuronalen Netz werden dabei über die Funktion f, welche dabei durch das neuronale Netz nachgebildet wird, neue Werte der Zeitreihe bestimmbar. Diese neuen Werte der Zeitreihe werden im Anschluß dem neuronalen Netz NN zugeführt, welches daraus wiederum durch Nachbildung der Funktion f neue Werte der Zeitreihe bestimmt. Dieser iterative Vorgang wird solange fortgesetzt, bis eine hinreichende Genauigkeit der zu bestimmenden Werte erreicht wurde.

Zur genauen Bestimmung fehlender Werte mit Hilfe der Monte-Carlo-Methode wird von folgenden Grundlagen ausgegangen. Es ist hier zu beachten, daß alle Lösungen die Form

$$\int h(u, m) P(u|m) du \qquad (9b)$$

aufweisen, wobei u den Satz von unbekannten Variablen und m den Satz von bekannten Variablen bedeutet. Ein Integral dieser Form kann beispielsweise gelöst werden, indem Zufallsproben der unbekannten Variablen gemäß $P(u|m)$ gezogen werden. Beispielsweise werden diese Proben mit $u^1, \ldots, u^s$ bezeichnet. Daraus ergibt sich folgender Zusammenhang für die Annäherung:

$$\int h(u,m)P(u|m)du \approx \frac{1}{S}\sum_{s=1}^{S} h(u^s,m). \qquad (9c)$$

Es ist zu beachten, daß in dieser Gleichung u den Wert y_{t-k}, welcher fehlt, entspricht. Mit dieser erfindungsgemäßen Lösung reduziert sich das Problem also darauf, aus $P(u|m)$ Proben zu ziehen. Für den Fall, daß lediglich eine Variable fehlt, also beispielsweise lediglich eine Rückführung aufgetrennt wurde, reduziert sich das Problem also auf das Probenziehen aus einer einvariablen Verteilung, welche mit Hilfe des "sampling-importance-resampling" oder anderen sampling-Techniken [1] getan werden kann.

Für den Fall, daß mehr als ein Meßwert fehlt, wird die Situation etwas komplizierter. Der Grund besteht darin, daß die unbekannten Variablen in der Regel voneinander abhängen und daß von der Verteilung aller unbekannten Variablen gezogen werden muß. Eine allgemeine Lösung dafür gibt das Gibbs-Sampling [1] an. Beim Gibbs-Sampling werden die unbekannten Variablen mit Zufallswerten, oder besser mit geschätzten Ausgangswerten initialisiert, welche beispielsweise aus den benachbarten Werten der fehlenden Werte abgeleitet werden können. Im Anschluß wird eine der unbekannten Variablen u_i ausgewählt und eine Probe von $P(u_i|m,u/u_i)$ gezogen; dann wird u_i auf diesen Wert gesetzt. Daraufhin wird die Prozedur für die nächste unbekannte Variable wiederholt usw. Abgesehen von beispielsweise den ersten Proben werden die Proben vorzugsweise mit der korrekten Fehlerverteilungsdichte gezogen. Dies bedeutet jedoch, daß für alle Unbekannten, die jemals in der Zeitreihe auftraten, Proben gezogen werden müssen. In der Praxis kann aber beispielsweise das Zeitfenster aus dem Proben gezogen werden auf eine vernünftige Größe beschränkt werden. Beispielsweise kann diese Größe der Größe des Markov blankets für die fehlenden Werte entsprechen. Hierbei ist zu beachten, daß für den Fall, daß zwischen zwei fehlenden Werten N aufeinanderfolgende Werte bekannt sind, die Kopplung zwischen den Unbekannten aufbricht und daß weitere Werte der Zeitreihe deshalb nicht berücksichtigt werden müssen.

Das Probenziehen für zukünftige Werte ist nach dem erfindungsgemäßen Verfahren besonders einfach. Es ist jedoch zu beachten, daß es so nicht für deterministische Systeme funktioniert. Durch die erfindungsgemäße Vorgehensweise wird eine besonders einfache Lösung für diese besonders kompliziert erscheinenden Sachverhalte gefunden. Beim Training des Netzes wird beispielsweise der Durchschnitt der Fehlergradienten gebildet, indem zu deren Berechnung die Werte der Zeitreihe verwendet werden, welche mit den Proben bestimmt wurden. Beim Ziehen der Proben nach der Monte-Carlo-Methode kann beispielsweise wie folgt vorgegangen werden.

Es sollen beispielsweise K-Schritte in der Zukunft der Zeitreihe vorhergesagt werden. Im Zusammenhang mit dem oben besprochenen Gleichungen bedeutet dies, daß die Werte $y_{t-1}, \ldots, y_{t-K+1}$ fehlen, und daß unter diesen Voraussetzungen y_t vorhergesagt werden soll. Unter diesen Voraussetzungen ist die Monte-Carlo-Methode sehr einfach. Zunächst muß beispielsweise von der Verteilung $P(y_{t-K+1}|y_{t-K}, \ldots, y_{t-K-N})$ eine Probe gezogen werden. Diese Probe wird mit y^s_{t-K+1} bezeichnet. Mit dieser Probe und den vorangegangenen Messungen wird beispielsweise eine weitere Probe y_{t-K+2} aus der Verteilung $P(y_{t-K+2}|y^s_{t-K+1}, \ldots, y_{t-K+1-N}$ usw. bis jede Probe für jede Unbekannte erzeugt wurde. Wenn man diese Prozedur S-mal wiederholt, so erhält man

$$E(y_t, M_{t-1}) = \frac{1}{S}\sum_{s=1}^{S} f(y^s_{t-2}, y^s_{t-1}, \ldots y^s_{t-N}) \qquad (9d)$$

Die experimentellen Befunde haben gezeigt, daß mit dem erfindungsgemäßen Verfahren gute Trainingsleistungen erzielt werden.

Literatur

[1] Bernardo, J. M., Smith, A. F. M. (1994) Bayesian Theory. Wiley & Sons.

[2] Buntine, W. L. and Weigend, A. S. (1991). Bayesian Back-Propagation. Complex systems, Vol. 5, pp 605—643.

[3] Ghahramani, Z. and Jordan, M. I. (1994. Supervised Learning from Incomplete Data via an EM approach. In: Cowan, J. D. et al., eds., Advances in Neural Information Processing Systems 6, Morgan Kaufman.

[4] Tresp, V., Ahmed, S. and Neuneier, R. (1994). Training Neural Networks with Deficient Data. In: Cowan, J.D. et al., eds., Advances in Neural Information Processing Systems 6, Morgan Kaufman.

Patentansprüche

1. Lernverfahren für ein rekurrentes neuronales Netz mit folgenden Merkmalen:
   a) mindestens eine von einem Ausgang zu einem Eingang zurückgeführte Verbindungsleitung des Netzes wird zum Lernen aufgetrennt und der entstehende Eingang in Form eines Lerneingangs

bereitgestellt

b) die Eingangsgrößen, welche dem Netz zum Training zugeführt werden, werden als Zeitreihe aufgefaßt;

c) von einem den Eingangsgrößen überlagerten unkorrelierten Rauschen endlicher Varianz, das im zeitlichen Mittel Null ist, wird die statistische Rauschverteilung ermittelt oder vorgegeben;

d) die dem Lerneingang zugeführte Eingangsgröße wird als fehlender Meßwert in der Zeitreihe in Form eines Fehlwertes aufbereitet, indem mindestens aus dem Fehlwert in der Zeitreihe benachbarten Meßwerten und der bekannten statistischen Rauschverteilung dessen statistische Fehlwert-Rauschverteilung berechnet wird und für den Fehlwert gemäß der Fehlwert-Rauschverteilung mindestens eine Monte-Carlo-Probe des Fehlwertes gezogen wird, die an seine Stelle in der Zeitreihe tritt.

2. Verfahren nach Anspruch 1, bei dem ein vorherzusagender Wert in der Zeitreihe als Prognosewert bestimmt wird, indem für den Fehlwert mehrere Monte-Carlo-Proben gezogen und deren Prognosewerte bestimmt werden, wobei als Prognosewert das arithmetische Mittel aus allen über die Proben bestimmten Prognosewerten verwendet wird.

3. Verfahren nach Anspruch 1 oder 2, bei dem mindestens zwei von einem Ausgang zu einem Eingang zurückgeführte Verbindungsleitungen des Netzes zum Lernen aufgetrennt werden und die entstehenden Eingänge in Form eines ersten und eines zweiten Lerneinganges bereitgestellt werden und bei dem von den auf diese Weise erzeugten zwei fehlenden und unmittelbar benachbarten Werten der Zeitreihe, welche dem ersten und zweiten Lerneingang beim Training zugeführt werden, zunächst der Erste aufbereitet wird und im Anschluß daran unter Zuhilfenahme des zuerst aufbereiteten der Zweite bestimmt wird.

4. Verfahren nach einem der vorangehenden Ansprüche, welches mehrfach ausgeführt wird.

5. Verfahren nach einem der vorangehenden Ansprüche, bei dem bei einem Lernschritt bei der Backpropagation die Lernschrittweite für die auf 1 normierten Eingangsgrößen des neuronalen Netzes zu 0,1 dividiert durch die Anzahl der gezogenen Proben festgelegt wird.

6. Verfahren nach einem der vorangehenden Ansprüche, bei dem als statistische Rauschverteilung eine Gaußverteilung festgelegt wird.

7. Verfahren nach einem der vorangehenden Ansprüche mit einer Zeitreihe der Form:

$$y_t = f(y_{t-1}, y_{t-2}, \dots, y_{t-N}) + \varepsilon_t \quad (1)$$

mit: ε_t: statistische Rauschverteilung
y: Meßwerte der Zeitreihe
y_t: vom neuronalen Netz vorherzusagender Wert
bei der die Funktion f durch das Neuronale Netz modelliert wird und die statistische Fehlerverteilungsdichte zu:

$$P_\varepsilon(y_{t-1} - f(y_{t-1}, y_{t-2}, \dots, y_{t-N})) = P(y_t | y_{t-1}, y_{t-2}, \dots, y_{t-N}) \quad (2)$$

bestimmt wird, woraus die Monte-Carlo-Proben $y_{t-k}^1, \dots, y_{t-k}^s$ gezogen werden und sich ein erwarteter vom neuronalen Netz vorherzusagender Wert mit den gezogenen Proben zu:

$$E(y_{t-k}, m) = \frac{1}{S} \sum_{s=1}^{S} f(y_{t-k}^{s}, m) \quad (3)$$

ergibt
mit: y_{t-k} fehlender Meßwert in der Zeitreihe
$k \leq N$
m: alle bekannten Meßwerte der Zeitreihe
S: Anzahl der Proben.

8. Verfahren nach einem der vorangehenden Ansprüche, bei dem das Neuronale Netz mit mindestens einem aufbereiteten Wert nach folgender Lernfunktion trainiert wird:

$$w_{neu} = w_{alt} + \eta \frac{\partial L}{\partial w}$$

mit: w: Neuronengewicht
L: logarithmische Wahrscheinlichkeitsfunktion
η: Lernfaktor
wobei gilt:

$$\frac{\partial L}{\partial w} = \frac{1}{S} \sum_{s=1}^{S} \sum_{l=N+1}^{t} \frac{\partial \log P^M(y_{l-1}^s, \dots, y_{l-N}^s)}{\partial w}$$

mit: NN_w: Werte der Funktion aus dem neuronalen Netz
wobei für y^s_l Meßwerte der Zeitreihe Verwendung finden und, falls ein Wert nicht vorhanden ist, aus der Wahrscheinlichkeitsverteilungsdichte:

$p^M(y^l|y^m)$

Monte-Carlo-Proben gezogen werden mit:
m: alle bekannten Meßwerte der Zeitreihe.

9. Verfahren nach Anspruch 6, bei dem das Neuronale Netz mit mindestens einem aufbereiteten Wert nach folgender Lernfunktion trainiert wird:

$$w_{neu} = w_{alt} + \eta \frac{\partial L}{\partial w}$$

mit: w: Neuronengewicht
L: logarithmische Wahrscheinlichkeitsfunktion
η: Lernfaktor
wobei gilt:

$$\frac{\partial L}{\partial w} = \frac{1}{S} \sum_{s=1}^{S} \sum_{l=N+1}^{t} y_l^s - NN_w(y_{l-1}^s, \ldots, y_{l-N}^s)) \frac{\partial NN_w(y_{l-1}^s, \ldots, y_{l-N}^s)}{\partial w}$$

mit: NN_w: Werte der Funktion aus dem neuronalen Netz
wobei für y^s_l Meßwerte der Zeitreihe Verwendung finden und, falls ein Wert nicht vorhanden ist, aus der Wahrscheinlichkeitsverteilungsdichte:

$p^M(y^l|y^m)$

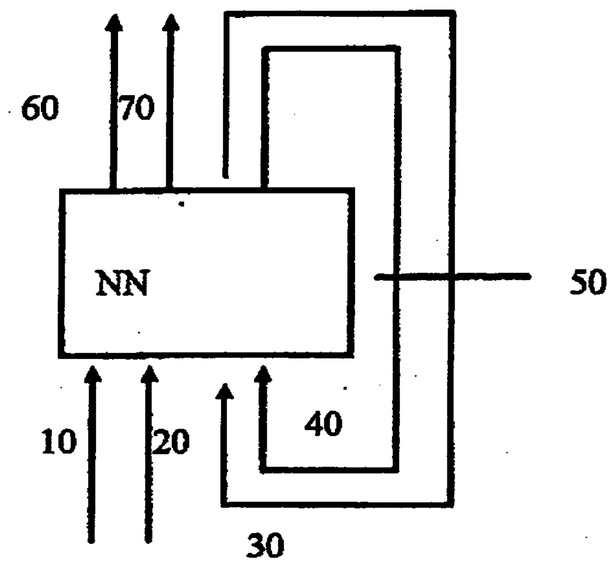
Monte-Carlo-Proben gezogen werden mit:
m: alle bekannten Meßwerte der Zeitreihe.

Hierzu 2 Seite(n) Zeichnungen

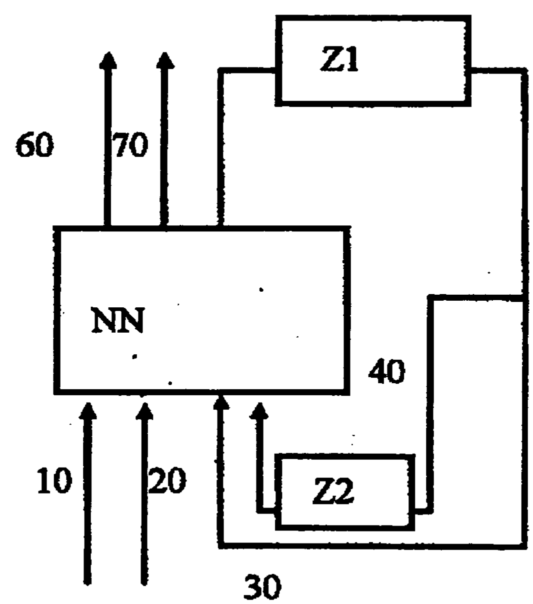
– Leerseite –

Nummer: DE 195 30 646 C1
Int. Cl.⁶: G 06 F 15/18
Veröffentlichungstag: 17. Oktober 1996

Figur 1

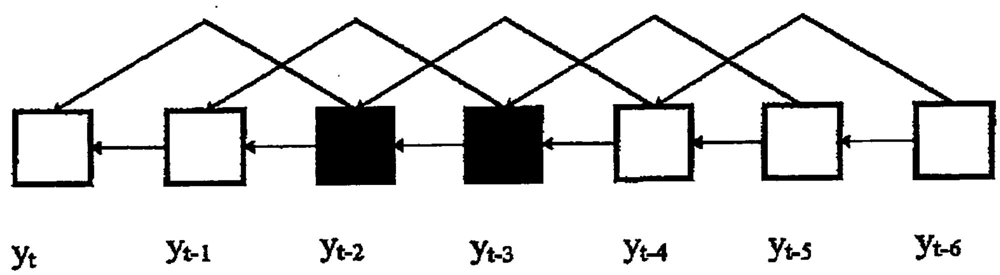


Figur 2



Nummer: DE 195 30 646 C1
Int. Cl.[6]: G 06 F 15/18
Veröffentlichungstag: 17. Oktober 1996

Figur 3



Figur 4